The InChI Chemical Structure Standard

Stephen Heller InChl-Trust Project Director

steve@inchi-trust.org

The main web sites for the IUPAC InChI project are:

http://www.iupac.org/inchi

and

http://www.inchi-trust.org

4/21/2016

Slides are available at http://www.hellers.com/steve/nanomolecules-4-16.pdf



This is a green talk –

These slides were made from 100% recycled electrons



InChl – The 30,000 foot view

This presentation will give the background and history of the InChI project.

The goal of this presentation is to show the value of InChI as a tool, for creating a notation for defined chemical structures and for linking to and finding more information than one can currently easily obtain.



www.inchi-trust.org

Make no little plans; they have no magic to stir men's blood and probably themselves will not be realized. Make big plans; aim high in hope and work.

~ Daniel Burnham



(With thanks to Francis Collins)



Date: Mon, 15 Nov 1999 18:48:30 -0500 (EST)

From: Stephen R. Heller<srheller@cliff.nal.usda.gov>

To: stein <sstein@enh.nist.gov> Subject: Re: A strawman proposal

Steve-

First rough draft. Let's talk tomorrow about it.

Steve

11/15/99

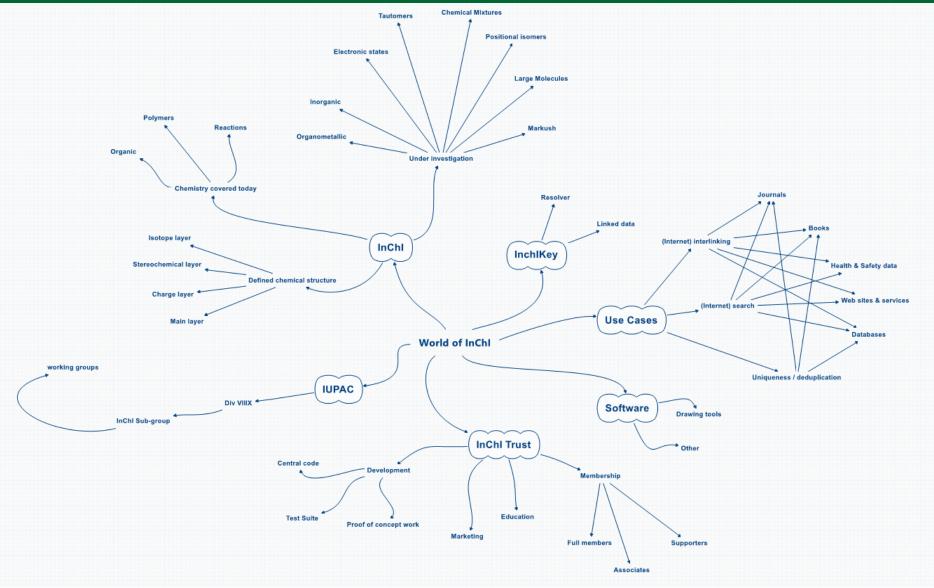
An IUPAC Chemical Registry System

In response to the upcoming March 2000 IUPAC meeting - Representations of Molecular Structure: Nomenclature and its Alternatives - I would like to propose the creation of an IUPAC public domain chemical registry system.

• • •



www.inchi-trust.org





Why InChl? - Too Many Good and Excellent Identifiers ("Standards")

Structure diagrams
- various conventions
- contain 'too much' information

Connection Tables/NotationsMolFiles, SDF, SMILES, SLN, ROSDAL, ...

Pronounceable names (and mostly unpronounceable) and mostly complex names - IUPAC, CAS 8th CI name, CAS 9th CI name, trivial, trade, WHO INN, ASK, ISO

(Dumb) Index Numbers
EINECS, ELINCS, FEMA, DOT, RTECS, CAS, Beilstein, USP, RTECS, EEC, RCRA, NCI, UN, USAN, EC, ChemSpider ID, REACH, PubChem CID, BAN, NSC, ASK, KEGG, BP, IND, MARTINDALE, MESH, IT IS, RX-CUI, NDF-RT, ATC, AHPA, USP/NF, UNII, MFCD#, ...



"Standards are like toothbrushes – everyone has one but no one wants to use someone else's."

Phil Bourne, Associate Director for Data Science, NIH



InChl is the worst computer readable structure representation except for all those other forms that have been tried from time to time.

With apologies to Sir Winston Churchill (House of Commons speech on Nov. 11, 1947)



InChl Project Goal:

To link data about a chemical from many sources with the purpose of creating new information.



Current InChl Status

At present, practically speaking, InChI can handle simple organic molecules, which turns out to cover 99%+ of what people deal with every day. If it did not then the usage of InChI would not be as great as it is.

InChI does not currently handle nanomolecules structures. To date no one or group has shown an interest in creating an approved IUPAC standard for nanomolecules.

InChI will handle nanomolecules. The issue is when, not if, and with whom InChI will work with to implement a nanomolecule standard.



What is InChl?

The IUPAC International Chemical Identifier, or InChl, is a non-proprietary, machine-readable string of symbols which enables a computer to represent the compound in a completely unequivocal manner.

InChls are produced by computer from structures drawn onscreen with existing structure drawing software, and the original structure can be regenerated from an InChl with existing structure drawing software.

InChI is really just a synonym.

http://en.wikipedia.org/wiki/International_Chemical_Identifier



InChl Videos

1. What on Earth is InChl?

http://www.youtube.com/watch?v=rAnJ5toz26c

2. The Birth of the InChl

http://www.youtube.com/watch?v=X9c0PHXPfso

3. The Googlable InChlKey

http://www.youtube.com/watch?v=UxSNOtv8Rjw

4. InChI and the Islands

http://www.youtube.com/watch?v=qrCqJ0o4jGs



InChl is plumbing. InChl is an (enabling) tool. InChl is a modern enabling technology.

For all but small group of chemists developing it, InChl is not something anyone should want to know about.

All you want to do is use InChI to find information on the web.

InChl is helping scientists to do better work and find/link to the latest information.



InChl is not a replacement for any existing internal structure representations. InChl is in ADDITION to what one uses internally. Its value to the is in FINDING and LINKING information



InChl is for computers

An InChI string is not directly intelligible to the normal human reader. Like Bar Codes, and InChI QR codes - InChIs are not designed to be read by humans.

Or, put another way – never send a human to do a machine's job!

Technology is at its best when it is invisible.



What "is" the InChl standard?

The InChI standard programmed into the algorithm is an arbitrary decision as to how structures are handled. In most cases there is total agreement (e.g., methane). In cases of more complex molecules where there is not agreement among chemists, one representation is chosen. As long as the arbitrarily chosen representation is properly programmed, one will always get the SAME result using it – which is what a standard is!



Unique InChl Features

Only IUPAC endorsed International structure standard

Only Open Source structure standard

Only structure standard support by a wide majority of publishers, database producers, and chemistry software companies



Whatever the controversies, InChI has now been more widely adopted than SMILES. In addition three US Government agencies - FDA, NIH, NIST - now have become paying members of the InChI Trust which would seem to indicate more official and institutional support leading to further widespread usage.



Large Databases with InChls/InChlKeys

NCI – 110 million
PubChem - 91 million (68 million online)
EBI UniChem – 90.5 million
ChemSpider – 34 million
IUPAC – 0 million



Why InChl is a success

- 1. Organizations need a structure representation for their content (databases, journals, chemicals for sale, products, and so on) so that their content can be LINKED to and combined with other content on the Internet. InChI provides an excellent ROI (return on investment). InChI increases productivity!
- 2. InChl is a public domain algorithm that anyone, anywhere can freely use. And they sure use it!

Success is uncoerced adoption



The main reason InChI works so well and at such a low cost is that I consider it a Crowdsourcing project.

The Trust gets the needed services (the creation of InChls) by the contributions from a large group of people rather than from traditional employees of an organization.



Four Requirements for a Computer Representation Standard

Need
Definition/Specification
Timing/Infrastructure
Acceptance/Use



An Open Source system keeps us on our toes. If things don't work or we don't respond as needed InChl won't remain a standard.



InChI is an international computer readable standard not just for chemists, but rather has very wide technical and non-technical use for linking and connecting information in many areas of scientific and everyday activities --

abstracting services biology/genomics databases bio-activity databases books chemical spills chemistry databases clinical trials company annual reports drug information drug overdoses electronic books environmental information food additives lawsuits magazines medical information medical records newspapers patents packages/bottles/transportation labels/ everyday product labels scientific journals toxicological information



Re: [CHMINF-L] Inchi and chemical databases

You forwarded this message on 9/15/2010 5:37 PM.

CHEMICAL INFORMATION SOURCES DISCUSSION LIST [CHMINF-L@LISTSERV.INDIANA.EDU] on behalf of Ian A Watson

Sent: Wednesday, September 15, 2010 3:24 PM

To: CHMINF-L@LISTSERV.INDIANA.EDU

Interesting example of Caffeine smiles on the web site. I was able to generate 172 different smiles for the Caffeine molecule (email me if you'd like them). Presumably each one of these could be a unique smiles in somebody's implementation.

But when I converted each of those 172 different smiles to InChI, I got the exact same InChI string for each one. That's exactly how things are supposed to work. Nice.

Ian Watson



c1(=0)c2c(n(C)c(=0)n1C)ncn2C c12c(n(C)c(=0)n(C)c1=0)ncn2C O=c1n(C)c(=O)c2c(ncn2C)n1C Cn1c2c(nc1)n(C)c(=0)n(C)c2=0 c12c(ncn1C)n(C)c(=0)n(c2=0)C O=c1c2c(ncn2C)n(c(=0)n1C)C c12c(n(cn1)C)c(=0)n(C)c(=0)n2C Cn1c2c(nc1)n(c(=0)n(c2=0)C)C c12c(ncn1C)n(c(=0)n(C)c2=0)C c12c(ncn1C)n(C)c(=0)n(C)c2=0 Cn1c(=0)n(C)c(=0)c2c1ncn2Cn1(c2c(nc1)n(C)c(=0)n(C)c2=0)C c12c(n(C)cn1)c(=0)n(c(=0)n2C)C Cn1c(=0)c2c(ncn2C)n(c1=0)C n1cn(C)c2c1n(c(=0)n(c2=0)C)C n1cn(c2c1n(C)c(=0)n(c2=0)C)C c12c(c(=0)n(c(=0)n1C)C)n(C)cn2 c1nc2c(n1C)c(=0)n(C)c(=0)n2C c1(=0)n(C)c(=0)c2c(ncn2C)n1C O=c1n(c(=O)c2c(ncn2C)n1C)C Cn1cnc2c1c(=0)n(C)c(=0)n2C n1(c(=0)n(c(=0)c2c1ncn2C)C)C c1(=0)n(C)c(=0)c2c(n1C)ncn2C O=c1n(c2c(n(cn2)C)c(=0)n1C)C Cn1c2c(n(cn2)C)c(=0)n(c1=0)CCn1c(=0)c2c(n(c1=0)C)ncn2C Cn1cnc2c1c(=0)n(c(=0)n2C)C c1nc2c(c(=0)n(C)c(=0)n2C)n1C c12c(ncn1C)n(c(=0)n(c2=0)C)C c1nc2c(n1C)c(=0)n(c(=0)n2C)C Cn1c2c(n(cn2)C)c(=0)n(C)c1=0 n1(C)c2c(n(C)c(=0)n(c2=0)C)nc1 n1(C)c2c(nc1)n(C)c(=0)n(c2=0)C n1(c(=0)c2c(n(c1=0)C)ncn2C)C n1(c(=0)c2c(n(C)c1=0)ncn2C)C Cn1c(=0)n(c2c(c1=0)n(C)cn2)C n1(C)c(=0)n(C)c(=0)c2c1ncn2C c1(=0)n(c(=0)c2c(ncn2C)n1C)C n1(cnc2c1c(=0)n(c(=0)n2C)C)C n1(C)c(=0)n(C)c2c(n(cn2)C)c1=0 n1(c2c(n(cn2)C)c(=0)n(C)c1=0)C n1(C)cnc2c1c(=0)n(C)c(=0)n2C O=c1c2c(n(C)c(=O)n1C)ncn2C n1(c2c(nc1)n(c(=0)n(c2=0)C)C)C n1(C)c(=0)c2c(n(c1=0)C)ncn2C n1(c2c(c(=0)n(C)c1=0)n(cn2)C)C c12c(n(c(=0)n(c1=0)C)C)ncn2C n1cn(C)c2c1n(C)c(=0)n(c2=0)C c12c(c(=0)n(C)c(=0)n1C)n(cn2)C Cn1c2c(n(C)cn2)c(=0)n(c1=0)Cn1(c(=0)n(C)c2c(n(cn2)C)c1=0)C n1cn(c2c1n(C)c(=0)n(C)c2=0)C c1(=0)n(c2c(c(=0)n1C)n(C)cn2)C Cn1c(=0)n(c(=0)c2c1ncn2C)CO=c1n(c(=O)n(c2c1n(cn2)C)C)C n1(c2c(c(=0)n(c1=0)C)n(C)cn2)C c12c(n(cn1)C)c(=O)n(c(=O)n2C)C c12c(c(=0)n(C)c(=0)n1C)n(C)cn2 Cn1c(=0)c2c(n(C)c1=0)ncn2C

```
c1(=0)n(C)c2c(n(cn2)C)c(=0)n1C
O=c1n(C)c2c(c(=O)n1C)n(C)cn2
n1(C)c2c(c(=0)n(C)c1=0)n(C)cn2
n1cn(c2c1n(c(=0)n(C)c2=0)C)C
O=c1n(c(=O)n(C)c2c1n(cn2)C)C
c1(=0)c2c(n(c(=0)n1C)C)ncn2C
c1(=0)n(c2c(n(cn2)C)c(=0)n1C)C
Cn1c2c(c(=0)n(c1=0)C)n(cn2)C
c1(=0)n(c(=0)c2c(n1C)ncn2C)C
O=c1n(c(=0)c2c(n1C)ncn2C)C
n1cn(C)c2c1n(c(=0)n(C)c2=0)C
n1(c(=0)n(C)c2c(c1=0)n(C)cn2)C
O=c1c2c(ncn2C)n(C)c(=O)n1C
n1(cnc2c1c(=0)n(C)c(=0)n2C)C
n1(C)cnc2c1c(=0)n(c(=0)n2C)C
n1cn(C)c2c1n(C)c(=0)n(C)c2=0
O=c1n(C)c(=O)n(C)c2c1n(C)cn2
n1(C)c(=0)n(c2c(c1=0)n(C)cn2)C
Cn1c(=0)c2c(ncn2C)n(C)c1=0
n1(c2c(n(cn2)C)c(=0)n(c1=0)C)C
n1(C)c2c(n(C)c(=0)n(C)c2=0)nc1
Cn1c2c(n(c(=0)n(c2=0)C)C)nc1
n1(c(=0)n(C)c(=0)c2c1ncn2C)C
O=c1n(C)c2c(n(C)cn2)c(=O)n1C
n1(C)c2c(n(cn2)C)c(=0)n(C)c1=0
c1(=0)c2c(ncn2C)n(c(=0)n1C)C
O=c1n(c2c(c(=O)n1C)n(cn2)C)C
Cn1c2c(n(C)c(=0)n(C)c2=0)nc1
Cn1c2c(nc1)n(c(=0)n(C)c2=0)C
Cn1c2c(n(C)cn2)c(=0)n(C)c1=0
c12c(n(C)c(=0)n(c1=0)C)ncn2C
n1(c2c(c(=0)n(c1=0)C)n(cn2)C)C
c1(=0)n(C)c(=0)n(c2c1n(cn2)C)C
n1(c2c(n(C)cn2)c(=0)n(c1=0)C)C
c1(=0)n(c2c(n(C)cn2)c(=0)n1C)C
n1(c2c(nc1)n(C)c(=0)n(c2=0)C)C
Cn1c2c(nc1)n(C)c(=0)n(c2=0)C
c12c(c(=0)n(c(=0)n1C)C)n(cn2)C
Cn1c2c(n(c(=0)n(C)c2=0)C)nc1
c1(=0)n(c(=0)n(C)c2c1n(C)cn2)C
c1(=0)n(C)c2c(n(C)cn2)c(=0)n1C
n1(c(=0)c2c(ncn2C)n(C)c1=0)C
n1(c2c(n(C)c(=0)n(C)c2=0)nc1)C
O=c1n(c2c(n(C)cn2)c(=O)n1C)C
c1(=0)n(C)c(=0)n(C)c2c1n(C)cn2
Cn1c(=0)n(c2c(c1=0)n(cn2)C)C
n1(c2c(n(c(=0)n(C)c2=0)C)nc1)C
Cn1c2c(c(=0)n(c1=0)C)n(C)cn2
c1(=0)n(C)c2c(c(=0)n1C)n(cn2)C
O=c1n(C)c2c(c(=O)n1C)n(cn2)C
c1(=0)n(C)c(=0)n(c2c1n(C)cn2)C
Cn1c(=0)n(C)c2c(n(C)cn2)c1=0
n1(c2c(nc1)n(c(=0)n(C)c2=0)C)C
O=c1n(c(=0)n(c2c1n(C)cn2)C)C
O=c1n(C)c(=O)n(C)c2c1n(cn2)C
c1(=0)n(C)c2c(c(=0)n1C)n(C)cn2
c1(=0)n(c(=0)n(C)c2c1n(cn2)C)C
n1(C)c(=0)c2c(ncn2C)n(C)c1=0
Cn1c(=0)n(c2c(n(C)cn2)c1=0)C
```

www.inchi-trust.org

O=c1c2c(n(c(=O)n1C)C)ncn2C O=c1n(C)c2c(n(cn2)C)c(=O)n1C n1(C)c(=0)n(c2c(n(C)cn2)c1=0)C n1(C)c2c(c(=0)n(c1=0)C)n(cn2)C Cn1c2c(c(=0)n(C)c1=0)n(C)cn2 c1(=0)n(c2c(c(=0)n1C)n(cn2)C)C n1(c2c(n(C)c(=0)n(c2=0)C)nc1)C n1(c2c(c(=0)n(C)c1=0)n(C)cn2)Cn1(C)c(=0)c2c(ncn2C)n(c1=0)C Cn1c(=0)n(C)c2c(n(cn2)C)c1=0 O=c1n(C)c(=O)c2c(n1C)ncn2C n1(c(=0)n(c2c(c1=0)n(cn2)C)C)C O=c1n(c(=O)n(C)c2c1n(C)cn2)C n1(C)c(=0)n(c2c(n(cn2)C)c1=0)C n1(c(=0)n(C)c2c(n(C)cn2)c1=0)Cc1(=0)n(C)c(=0)n(C)c2c1n(cn2)C n1(c(=0)n(C)c2c(c1=0)n(cn2)C)C O=c1n(C)c(=O)n(c2c1n(cn2)C)C n1(c(=0)c2c(ncn2C)n(c1=0)C)C c1(=0)c2c(ncn2C)n(C)c(=0)n1C Cn1c2c(n(C)c(=0)n(c2=0)C)nc1 n1(C)c(=0)c2c(n(C)c1=0)ncn2C n1(C)c(=0)n(C)c2c(c1=0)n(C)cn2 Cn1c2c(c(=0)n(C)c1=0)n(cn2)C n1(C)c(=0)n(C)c2c(n(C)cn2)c1=0 n1(c2c(n(C)cn2)c(=0)n(C)c1=0)Cn1(C)c(=0)n(c(=0)c2c1ncn2C)C c1(=0)n(c(=0)n(c2c1n(cn2)C)C)C c1(=0)n(c(=0)n(c2c1n(C)cn2)C)C n1(C)c2c(nc1)n(c(=0)n(C)c2=0)C Cn1c(=0)n(C)c2c(c1=0)n(C)cn2 O=c1n(c2c(c(=O)n1C)n(C)cn2)C n1(C)c2c(n(c(=0)n(c2=0)C)C)nc1 n1(C)c(=0)n(C)c2c(c1=0)n(cn2)C n1(C)c2c(nc1)n(C)c(=0)n(C)c2=0 n1(C)c2c(n(cn2)C)c(=0)n(c1=0)C n1(C)c(=0)n(c2c(c1=0)n(cn2)C)Cn1(C)c2c(c(=0)n(C)c1=0)n(cn2)C n1(c(=0)n(c2c(n(C)cn2)c1=0)C)Cn1(c(=0)n(c2c(c1=0)n(C)cn2)C)C n1(C)c2c(n(C)cn2)c(=0)n(C)c1=0 n1(C)c2c(c(=0)n(c1=0)C)n(C)cn2

n1(C)c2c(n(c(=0)n(C)c2=0)C)nc1

n1(C)c2c(nc1)n(c(=0)n(c2=0)C)C

www.inchi-trust.org



InChl

172 SMILES representations





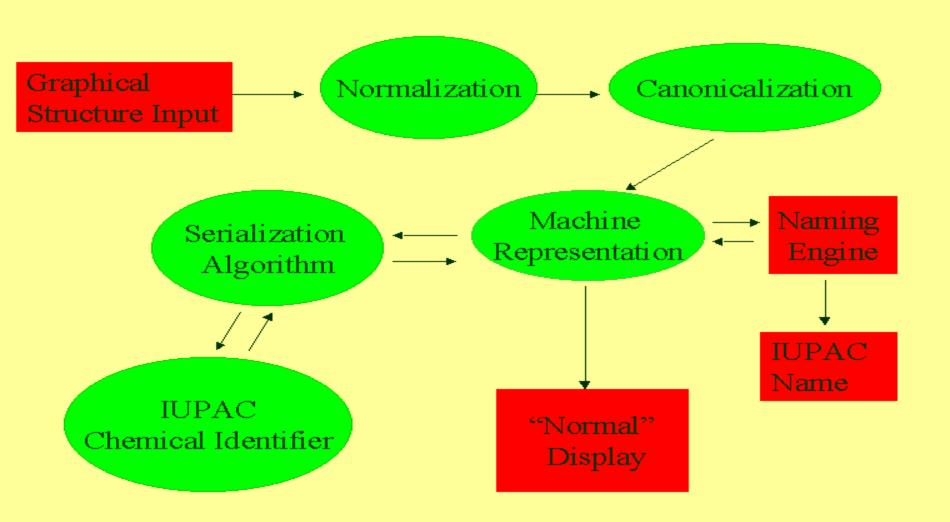
E Pluribus Unum Out of many, One



InChl Characteristics

- 1. Easy to generate (It will use existing software.)
- 2. Expressive (It will contain structural information.)
- 3. Unique/Unambiguous
- 4. Easy to search for structure via Internet search engines (Google, Yahoo, Bing, etc.) using the InChl (hash) Key.







InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H.1-3H3 (caffeine)

character indicating the number of protons ('N' means neutral)

InChikev=RYYVLZVUVIJVGH-UHFFFAOYSA-N

ev=RYYVLZV

First block (14 letters)

Encodes molecular skeleton (connectivity) Second block (8 letters)

Encodes stereochemistry and isotopes

flag character for InChI version: 'A' for version 1

flag character ('S') indicates standard InChlKey (produced out of standard InChl)



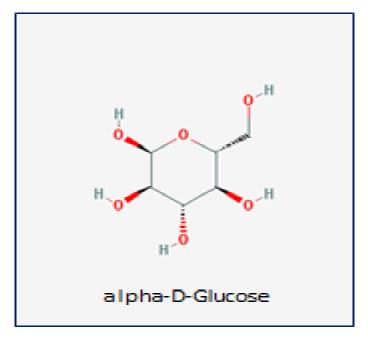
InChl is a string

InChI=1S/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1H2/t2-

,3-,4+,5-,6+/m1/s1

Version/Type
Chemical formula
Connectivity
Charge/Proton
Stereochemical
Other (e.g., Isotopic)

"layered" line notation





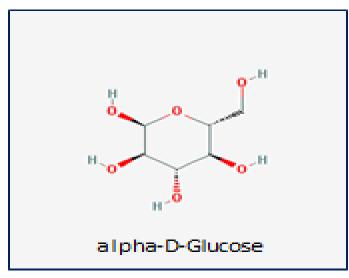
InChlKey is a "hashed" InChl

- Search engine friendly InChl
- May allow for 'secure' lookup of a chemical

WQZGKKKJIJFFOK-DVKNGEFBSA-N

Chemical formula
Connectivity
Stereochemical
Other (e.g., Isotopic)
Type
Version
Charge/Proton

"layered" line notation





InChI for Maitotoxin (from Nextmove Software, UK)

www.inchi-trust.org



What about funding?



Don't give up - Moses was once a basket case



While InChI did not make the top 10 given to Moses, it is #14

(Thou shall use InChl for structure representation.)



The InChl Trust

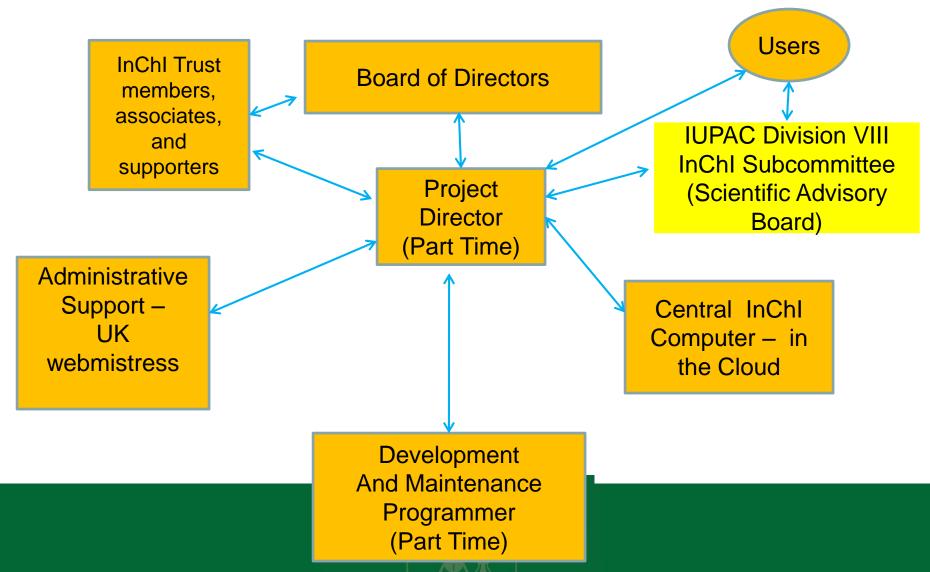


The InChl Trust

To function and succeed, InChI had to become personality independent. InChI had to be "institutionalized". If the work of this project was to be enduring it needed to turned over to an entity that would ensure its ongoing activities and be acceptable to the community. It was concluded that a not-for-profit organization would best fit the ongoing and future project needs. Thus the decision to create and incorporate the "InChI Trust" as a UK charity.



InChl Trust Organization



InChl Staff and Collaborators

The InChI project has had the unusual perfect "good storm" of cooperation and support. It is a truly international project with programming in Moscow, computers in the cloud, incorporated in the UK, and a project director in the USA. Collaborators from over a dozen countries, from academia, Pharma, publishers, and the chemical information industry, have all offered, and continue to offer, senior scientific staff to develop the InChI standard.



The Future

InChl has become mainstream for publishers, databases providers, and software developers. Over the next 5-10 years, publishers will use data mining to create both better abstracts, useful indexing, and concept terms. Search engines will be able to search for appropriate text and structures and direct users to the original (fee or free/Open Access/Open Data) sources.



Summary

If you are not part of the solution; you are part of the precipitate



Acknowledgements

(Primarily members for the IUPAC InChI subcommittee and associated InChI working groups)

Steve Bachrach, Colin Batchelor, John Barnard, Evan Bolton, Steve Boyer, Steve Bryant, Szabolcs Csepregi, Rene Deplanque, Gary Mallard, Nicko Goncharoff, Jonathan Goodman, Guenter Grethe, Richard Hartshorn, Jaroslav Kahovec, Richard Kidd, Hans Kraut, Alexander Lawson, Peter Linstrom, Bill Milne, Gerry Moss, Peter Murray-Rust, Heike Nau, Marc Nicklaus, Carmen Nitsche, Matthias Nolte, Igor Pletnev, Josep Prous, Peter Murray-Rust, Hinnerk Rey, Ulrich Roessler, Roger Schenck, Martin Schmidt, Steve Stein, Peter Shepherd, Markus Sitzmann, Chris Steinbeck, Keith Taylor, Dmitrii Tchekhovskoi, Bill Town, Wendy Warr, Jason Wilde, Tony Williams, Andrey Yerin.

Special Acknowledgement: Ted Becker & Alan McNaught for their vision and leadership of the future of IUPAC nomenclature.

